What Teachers and Students Know about Data Management

Andreas Grillenberger and Ralf Romeike

Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany
{andreas.grillenberger, ralf.romeike}@fau.de

Abstract. Data management is a highly innovative field of CS, which evolved from the original field databases in the last years. With the ongoing developments, several topics from this field, such as cloud computing, large data collections or data analyses, pervade our daily lives. Although more and more students and teachers come in contact with data management topics and need to develop competencies in this field, current CS education typically does not sufficiently address them. Yet, both students and teachers already have experience with certain aspects of data management and may have built up knowledge and perceptions, which need to be considered in CS teaching. Hence, in a qualitative study, we investigated the attitudes and prior knowledge of teachers on several data management topics and explored students' knowledge in this field.

Keywords: Data Management, Knowledge, Experiences, Teachers, Students

1 Introduction

In recent years, new requirements and technologies led to the formation of *data management* as a new field of CS, in particular due to continuously increasing amounts of data being stored and analyzed. Although it is highly relevant in CS today and becomes increasingly pervasive in everyone's daily lives, secondary CS education sets its focus predominantly on other fields of CS. Nowadays, in lessons on data-oriented topics, there is a clear emphasis on databases and database-related aspects, while other parts of data management are typically left out [6]. Considering data management topics in CS education can enrich current teaching and opens up various new possibilities, in particular because they are not only interesting from a scientific perspective, but also exemplary for the ongoing developments in CS. They also support the development of competencies that everyone needs for responsibly handling their own and others' personal data [7]. Consequently, more and more curricula and educational standards introduce topics such as data analysis, security and privacy (cf. e. g. [3]). Our experience shows that students and teachers are generally interested in topics related to big data and data management and that they regard competencies in this field as essential.

There is strong agreement that, besides central principles, also the prior knowledge of teachers and students should be considered when bringing new topics to school

(cf. e. g. [5]). Hence, we describe two investigations: First, we examine teachers' content knowledge about and attitudes towards typical data management topics, as well as the challenges they see for teaching. Second, in order to gain insight into students' experience, we describe an exploratory analysis of their knowledge in this field.

2 Data Management from a CS Perspective

In the last 10 to 15 years, *data management* has evolved from the field databases. A central topic in this field is *big data*, which deals with storing and analyzing large amounts of highly varied data as fast as possible (cf. e. g. [8]). With the increasing relevance of correlation-based data analyses ("data mining"), new requirements are imposed on data management systems, for example the need to store data distributed on multiple servers because of the high volumes. At the same time, ensuring a high velocity requires minimizing the amount of communication between the servers involved. Hence, various new systems and technologies have emerged and became important areas of data management research, for example non-relational *NoSQL databases*, *in-memory databases* or *cloud computing*. Correspondingly, new and highly innovative methods, approaches and principles were developed and became important to CS. Aside its relevance in CS research, the significance of data management topics in our daily lives has also massively increased: Nowadays, everyone uses various technologies based on data management techniques, comes into contact with metadata, stores data, protects and shares it and reads news about data-related topics, such as extensive data analyses by companies or intelligence agencies.

3 Related Work

Despite the significant developments in data management, hardly any research in this field, but also related to databases and data in general has been conducted in CS education research since database teaching was established in the early 1990s. Even in recent years, only occasional approaches were described, e. g. on introducing big data at high school [2]. In a qualitative analysis of curricula and teaching standards, we identified the gap between current CS education and the scientific perspective on this field [6], which in particular affects newer aspects of data management. Also, we already identified several key competencies everyone needs for handling data in everyday life, for example that students need to "*understand the consequences of synchronizing data and deal with synchronization conflicts*" [7].

Despite the importance of this field, students' and teachers' knowledge about and attitudes towards data management and the traditional topic databases, have not been investigated yet. Typically, research concerning students' perspectives assesses their preconceptions (e. g. [5]): For example, Diethelm et al. [4] presented an approach for identifying contexts relevant for the students by using the miracle question method. In studies on teachers' perspectives, their knowledge and attitudes are often considered by investigating their content knowledge or pedagogical content knowledge, which, to-

gether with the general pedagogical knowledge, are central for teaching [10]. For example, in the context of developing teacher training, Mesaroş & Diethelm [9] surveyed teachers in order to discover their ideas about lesson planning on specific topics.

4 Teachers' Content Knowledge and Attitudes

4.1 Aims

The investigation of teachers' perspectives on data management has various possible foci, e. g. their motivation, attitudes, content and/or pedagogical content knowledge or their experience with these topics. As data management is rather new to CS education, we expect that they have no teaching experience yet. Hence, in this study we concentrate on the following questions:

- What do teachers know about data management topics (content knowledge)?
- Which topics do they consider interesting for their teaching?
- Which challenges do they expect when including data management topics in their teaching?

Based on previous feedback from teachers, our hypothesis is that they have only limited knowledge about data management, except for traditional aspects such as databases and data modeling. Additionally, we assume that the complexity of the topics and a lack of suitable software could be seen as significant obstacles for CS teaching in this field.

4.2 Survey Method and Implementation

For investigating these questions, we surveyed 53 teachers prior to three teacher training workshops using questionnaires. We decided for this method, because the goal was not to get deep insight; instead, we wanted to get an overview of the teachers' knowledge, interest and expected challenges when including this topic in teaching. The participants were from three German federal states (36 from Bavaria, 17 from the Berlin/Brandenburg area) and different types of secondary schools. Among the teachers, 31 were master teachers in CS at their respective schools.

In the questionnaires, we presented the teachers a list of data management topics, which were selected in an empirical analysis of widely accepted literature on data management in previous work. This list is shown along with the results in Table 1. On each of these topics, we asked the teachers the following questions:

1. How do you rate your knowledge about each of the topics?
   four point Likert scale from "unknown'' to "detailed knowledge"
2. How interesting do you consider each topic for your teaching?
   four point Likert scale from "not interesting" to "very interesting"
3. Which challenges do you expect in lessons on this topic?
   options: insufficient own knowledge, missing tools, topic is too complex

4.3 Results and Interpretation

Before analyzing the results, we cleaned the data: When the answer to the first question stated that the topic is unknown, the answers to the other questions were not considered, as answering these is not possible without knowing about the topic. The data was then aggregated by calculating median and mode measures as well as the mean deviation from the median (MD) for every question and topic. As the dimensions of our questions are on an ordinal scale, these measures are appropriate for aggregating the data. The complete results are shown in Table 1.

In general, the teachers state to have limited knowledge about the presented data management topics, but have already heard of most of them. This is the case even for teachers who consider data management topics interesting. Despite this, they estimated their knowledge about *relational databases* as rather detailed, while about other topics that are already considered in school, such as *data analysis*, *data encryption* or *metadata*, they supposed to have basic knowledge. For topics that are typically left out in current CS education, such as *distributed databases*, *big data* and *data mining*, they stated to have only little knowledge. One exception is *cloud storage*, on which they estimate their knowledge as basic. Merely three topics were unknown to them: the *CAP theorem*, the *ACID* and *BASE paradigms*[1]. For 13 of the 19 topics, the MD is below or at 0.25, while for all others it is at least below 0.5. Hence, most results show a relatively high consensus among the participants.

While the teachers rate *data security*, *data privacy* and *threats of automatic data processing* as very interesting for their teaching, they consider rather technology-oriented terms such as *non-relational* and *distributed databases*, *open data* or the underlying principles as less interesting. Yet, in general, most data management topics were rated rather interesting for teaching. Most of the answers have a high MD and hence a wide spread in the answers: with a closer look at the results, it becomes clear that most topics were rated as being very interesting by several teachers and at the same time as hardly interesting by others.

When including data management topics in their teaching, most teachers see the primary challenge in their insufficient knowledge. This and the results from the first question show a strong need for materials and teacher training that helps to build up this knowledge. In addition, teachers also see a challenge in missing tools that are suitable for CS teaching. Yet, in general, they do not expect to encounter any problems with the complexity of these topics, which may be influenced by their limited knowledge.

Resulting from these data, we can assume that although the participants of the workshops were generally interested in data management, they have only limited knowledge in this field. This is the case even for teachers who are master teachers at their respective schools. Hence, our results show a clear need for further education of CS teachers in data management topics.

---

[1] The *ACID paradigm* describes the four central characteristics of traditional databases, **a**tomicity, **c**onsistency, **i**solation and **d**urability. The *BASE paradigm* is central to non-relational databases, which are **b**asically **a**vailable, **s**oft-state, **e**ventually consistent. The *CAP theorem* concludes, that **c**onsistency, **a**vailability and **p**artition tolernace, cannot be achieved at the same time in a data management system [1].

Table 1. Results of the teacher questionnaire

	Knowledge 0=unknown 1=little knowledge 2=basic knowledge 3=detailed knowledge				Estimated interest 0 = not interesting 1 = hardly interesting 2 = rather interesting 3 = very interesting				Challenges % of teachers			
	# answers	mode	median	MD	# answers	mode	median	MD	insufficient knowledge	missing tools	com-plexity
relational databases	53	3	3	0.43	53	3	2	0.58	1.9	11.3	7.5
NoSQL / non-relational databases	52	1	1	0.02	33	1	1	0.39	49.1	11.3	11.3
distributed databases	52	1	1	0.15	34	1	1	0.50	41.5	15.1	11.3
cloud storage	53	2	2	0.11	49	2	2	0.12	30.2	20.8	1.9
cloud computing	51	1	1	0.45	44	2	2	0.02	35.8	28.3	11.3
data analysis	51	2	2	0.25	44	2	2	0.14	18.9	24.5	5.7
data mining	52	1	1	0.10	37	2	2	0.16	49.1	18.9	7.5
big data	51	1	1	0.27	38	2	2	0.16	39.6	24.5	9.4
open data	51	1	1	0.22	23	1	2	0.74	37.7	11.3	1.9
data encryption	53	2	2	0.11	50	3	3	0.50	13.2	18.9	11.3
data modeling	52	2	2	0.23	49	3	2	0.69	11.3	3.8	3.8
function of search engines	51	2	2	0.31	45	2	2	0.20	18.9	17.0	1.9
CAP theorem	51	0	0	0.14	3	1	1	0.00	47.2	0.0	0.0
ACID paradigm	51	0	0	0.39	6	1	2	0.50	37.7	0.0	0.0
BASE paradigm	51	0	0	0.12	4	1	1	0.00	45.3	0.0	0.0
meta data	51	1	1	0.45	36	2	2	0.22	32.1	17.0	3.8
data security (e. g. backup)	52	2	2	0.15	46	3	2	0.74	11.3	9.4	0.0
data privacy	53	2	2	0.21	50	3	3	0.46	11.3	17.0	1.9
threats of automatic data processing	51	2	2	0.02	48	3	3	0.60	7.5	22.6	3.8

5 Students' Knowledge and Experience

5.1 Aims

For school teaching, the students' prior knowledge about and experience with a topic are an important basis to build upon. Although data management is hardly represented in CS lessons, it is reasonable to assume that due to the ubiquity of these topics (and related technologies), students acquire some knowledge and gain experience e. g. through using smartphones and the Internet or by managing their personal data on computers. Exploring their knowledge in this field can thus help to get insight into how students come into contact with data management topics. Hence, our main question for the investigation is: *What do students know about specific topics of data management?*

5.2 Survey Method and Implementation

For exploring their knowledge, we surveyed 42 Bavarian students using questionnaires in extra-curricular settings. Among them, 38 are from higher secondary schools ("Gymnasium") and four from an intermediate secondary school ("Realschule"). Most students already came into contact with relational databases and data modeling. Yet, in school teaching other aspects of data management have hardly been considered. To explore the students' knowledge about data management, we asked them questions on:

1. Which knowledge do students have concerning the purpose and use of databases and data analyses?
2. Which metadata do students expect to be captured in situations from their daily life (taking photos with the smartphone, surfing the web)?
3. Which data do students estimate as valuable enough to create backups? How do they create backups?

The topics of the first question were selected because they are on the one hand central to data management, but also typical topics of secondary CS teaching (databases) or at least strongly related to current teaching (data analyses). Hence, we expected them to have at least little knowledge about these topics. In order to assess this knowledge, we presented them several statements (e. g. *"In databases, all data must be stored consistently"*, *"Metadata is often more interesting than the original data"*), for which they should decide whether they are correct or not. For the second question type, two situations related to the production and use of metadata were described (taking a photo with a smartphone, surfing the web), for which they should decide which metadata from a given list are stored/transmitted along with the original data. While in the first situation metadata is fairly obvious to students, in the second case students probably have not come into contact with it. Thus, the questions can give insight into whether the students are aware of metadata being stored, about what kind of information they think can be transmitted, and about their estimation of the extent of such data. The third question type refers to one exemplary topic of data management strongly related to the students' daily life and gives insight into how valuable data is for the students: The creation of

backups requires various considerations, for example selecting appropriate backup media (e. g. by creating backups on external drives or thumb drives or synchronizing their data with the cloud), deciding whether full or incremental backups are to be created, how long data is being stored and also which data to backup.

5.3 Results and Interpretation

All statements that the students could tick were treated as sub questions, for which the number of students who checked them was counted. The question on backups was treated separatel, since it was the only one with free text answers: For this, we extracted all responses and counted the respective number of answers. The results of all questions are shown in Table 2.

The results from the first question show that students have very vague knowledge concerning databases and data analyses. Although most of them have already attended lessons on databases, there is hardly any difference between the answers on questions related to typical topics of teaching and such that typically cannot be answered with school knowledge. About 38% of the students know that data analyses may be used for finding additional information that is not obviously contained in the original data, nearly 55% know that metadata is often more interesting than the original data. This suggests that they have already heard of these topics in daily life, e. g. in news reports on analyses of shopping habits at large online shops. About 45% of the students support the statement that small amounts of data should be preferred for analysis purposes, because analyzing them is faster, which suggests that their knowledge about data analyses is only superficial. In general, the results indicate that the participants of the survey have a basic but also very diverse knowledge about data analyses.

The second question give an impression about students' knowledge about metadata in two different contexts: Most of them know that date and time are captured when taking a photo with their smartphone. Also, about 60% know that the GPS location are stored along with the picture, as well as information about the camera/phone with which the photo was taken. Most students were correct in assuming that names of persons or a description of the photo are typically not stored automatically. On the second situation we described to them, surfing the web, a majority of the students assumed that the web server gets to know the client's web browser, about 78% expect that the user's country can be discovered and only 52% think the same applies to language and operating system. While several students underestimate the amount of metadata and thus do not expect the programs installed or the screen resolution to be disclosed, others overestimate the possibilities and even assume that web sites automatically get the user's mail address or information their interests.

These results show that they are generally aware that additional data may be collected when using devices like their smartphone, which is explicable as they encounter such data regularly. Yet, when metadata are created rather in the background, despite knowing the basic concept, fewer students are aware of the creation of such data and can estimate their extent. These results are in particular interesting for CS teaching, as it clearly shows that the students relate to the topic metadata in their daily life and have built up knowledge prior to data management lessons.

The question about backups gives insight into how important personal data is for students and how they protect it. Nearly 88% stated that they create backups regularly: 74% use external media such as thumb drives for this purpose, while 38.1% synchronize data to the cloud and about 29% use both methods. The others, nearly 17%, do not create any backups. Among those, the majority thinks that their data is not valuable enough, while three students stated that they did not even think about creating backups. Thus, in general the results show that their data is valuable for students and hence they want to protect it. The most important data is photos (52%) and videos (33%), followed by documents (14%).

Summarizing, the results show that students have already heard of several aspects of databases, data analyses and metadata. They are using at least two different approaches for data backup and probably take advantage of metadata stored along with photos. This confirms our hypothesis that they have prior knowledge about data management topics that should not be neglected when planning lessons. Particular topics on which they have wrong or incomplete conceptions, need to be addressed in data management teaching in order to foster a deeper understanding of such topics that are strongly related to their daily life.

6 Conclusion

Our teacher questionnaires clearly show that professional development opportunities on data management topics should be provided for teachers: Although they show significant interest and in some cases tried to incorporate data management topics in their teaching, they generally consider their own knowledge as insufficient. Thus, continuous professional development is deemed an important task. Despite their lack of knowledge, the participants do not expect that data management topics are too complex for secondary school teaching. Prior to the teacher training workshops, several teachers told us that they could not grasp this large field, as it included too many aspects that are unknown to them. Nevertheless, they also see the topics as motivating and interesting for themselves and their students. The discussions following the subsequent workshop have reinforced this impression.

While students come in contact with metadata of photos and thus know about them, they are not aware of which data is disclosed when surfing the web. Also, they have strategies concerning how to store and backup their own data. So, concerning the students' results, we conclude that there is rudimentary knowledge about data management topics, which teaching could build upon. Yet, they only have a vague understanding of the use and possibilities of data management topics. Generally-spoken, their knowledge is not sufficient for recognizing the ubiquity of data management and in particular for understanding influences on their daily lives. Students regularly encounter phenomena related to data management, for example when synchronization errors occur while using cloud storage services. However, their knowledge is typically not sufficient for understanding the reasons of these problems, for preventing them, and for deciding how to solve such conflicts.

Table 2. Results of the student questionnaire

		# answers
Q1:	**databases & data analyses**	
1.1	In databases, all data must be stored in a consistent way	12 (28.6%)
1.2	Only 5 users can use a database at the same time	1 (2.4%)
1.3	Each database is stored on an own server	9 (21.4%)
1.4	Cloud services typically use databases	25 (59.5%)
1.5	Data analyses always last very long	6 (14.3%)
1.6	Small amounts of data should be preferred as analyzing them is faster	19 (45.2%)
1.7	When analyzing large amounts of data, only few information can be found	6 (14.3%)
1.8	By data analyses, it is possible to find more information on users than contained in the original data	16 (38.1%)
1.9	Large amounts of data can hardly be analyzed	6 (14.3%)
1.10	Meta data are often more interesting than the original data	23 (54.8%)
Q2	**meta data of smartphone photos**	
2.1	date/time	41 (97.6%)
2.2	GPS location	25 (59.5%)
2.3	names of persons shown on the photo	3 (7.1%)
2.4	description of the photo	3 (7.1%)
2.5	name of the photographer	3 (7.1%)
2.6	information on the camera	25 (59.5%)
Q3	**meta data when accessing web sites**	
3.1	referring URL	21 (50%)
3.2	browser name	32 (76.2%)
3.3	operating system	22 (52.4%)
3.4	GPS location	15 (35.7%)
3.5	name of user	7 (16.7%)
3.6	names of several installed programs	5 (11.9%)
3.7	mail address of user	10 (23.8%)
3.8	interests of user	13 (31%)
3.9	unique user ID	8 (19%)
3.10	stationary or mobile device	26 (61.9%)
3.11	screen resolution	3 (7.1%)
3.12	language	22 (52.4%)
3.13	country	33 (78.6%)
3.14	age of user	2 (4.8%)
Q4	**backup**	
4.1	regular creation of backups on e. g. thumb drives	31 (73.8%)
4.2	synchronization with cloud	16 (38.1%)
4.3	my data are not valuable enough	7 (16.7%)
4.4	did not yet think about that	3 (7.1%)
4.5	data being backuped:	
4.5.1	photos	22 (52.4%)
4.5.2	documents	6 (14.3%)
4.5.3	videos	14 (33.3%)
4.5.4	school-related files	2 (4.8%)
4.5.5	savegames	1 (2.4%)
4.5.6	applications	1 (2.4%)
4.5.7	application data	3 (7.1%)
4.5.8	music	4 (9.5%)
4.5.9	contacts	4 (9.5%)

For supporting students' understanding of phenomena and consequences related to data management, CS education needs to further emphasize this field. In addition to basic knowledge about the concepts and principles, competencies need to be fostered that are necessary for understanding the public discourse on topics such as data storage and data analyses, for estimating and circumventing threats as well as for self-determined and responsible handling their own and others' personal data. On the other hand, the teachers' results emphasize the need for professional development opportunities and show clear starting points for developing appropriate materials in this field. In conclusion, for giving further guidelines on how to bring CS education to school, the results shown in this paper are a clear basis, but there is a strong need for further research.

References

1. Brewer, E.: CAP twelve years later: How the "rules" have changed. Computer 45(2), 23–29 (2012).
2. Buffum, P. S., Martinez-Arocho, A. G., Frankosky, M. H., Rodriguez, F. J., Wiebe, E. N., Boyer, K. E.: CS Principles goes to Middle School. In: Dougherty, J. D., Nagel, K. (eds.) Proceedings of the 45th ACM technical symposium on Computer science education, pp. 151-156. ACM, New York, NY, USA (2014).
3. CSTA Standards Taskforce: (Interim) CSTA K-12 Computer Science Standards. CSTA/ACM, New York, NY, USA (2016).
4. Diethelm, I., Borowski, C., Weber, T.: Identifying relevant CS contexts using the miracle question. In: Schulte, C., Suhonen, J. (eds.) Proceedings of the 10th Koli Calling International Conference on Computing Education Research, pp. 74-75. ACM, New York, NY, USA (2010).
5. Diethelm, I., Hubwieser, P., Klaus, R.: Students, teachers and phenomena: Educational reconstruction for computer science education. In: Laakso, M.-J., McCartney, R. (eds.) Proceedings of the 12th Koli Calling International Conference on Computing Education Research, pp. 164-173. ACM, New York, NY, USA (2012).
6. Grillenberger, A., Romeike, R.: A Comparison of the Field Data Management and its Representation in Secondary CS Curricula. In: Schulte, C., Caspersen, M. E., Gal-Ezer, J. (eds.) Proceedings of the 9th Workshop in Primary and Secondary Computing Education, pp. 29-36. ACM, New York, NY, USA (2014).
7. Grillenberger, A., Romeike, R.: Teaching data management: Key competencies and opportunities. In: Brinda, T., Reynolds, N., Romeike, R., Schwill, A. (eds.) KEYCIT 2014: Key Competencies in Informatics and ICT, pp. 133-150. Universitätsverlag Potsdam, Potsdam, Germany (2014).
8. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group, Stamford, CT, USA (2001).
9. Mesaroş, A.-M., Diethelm, I.: Ways of planning lessons on the topic of networks and the internet. In: Knobelsdorf, M., Romeike, R. (eds.) Proceedings of the 7th Workshop in Primary and Secondary Computing Education, pp. 70-73. ACM, New York, NY, USA (2012).
10. Shulman, L. S.: Those who understand: Knowledge growth in teaching. Educational Researcher 15(2), 4–14 (1986).